Proyecto Final - Curso R - Redbioma

Relación de area foliar con radiación solar en especies del género *Dipteryx*

Informe preparado por Jose Pablo Jimenez Madrigal

Introducción

El género *Dipteryx*, perteneciente a la familia Fabaceae, es de gran importancia ecológica y económica en las regiones tropicales de América del Sur y Central. Los árboles de *Dipteryx* son fundamentales para la conservación de la biodiversidad, ya que proporcionan hábitat y alimento a diversas especies de fauna silvestre. Su madera, dura y resistente, es también apreciada en la construcción y la ebanistería, lo que resalta su relevancia en la economía local. Más aún, varias especies del género producen frutos ricos en coumarina, un compuestos orgánico con aplicaciones en la industria farmacológica, cosmética y alimenticia . Por estas razones, la conservación de las especies del género *Dipteryx* es vital tanto para los ecosistemas donde se encuentran como para las comunidades humanas que dependen de sus múltiples beneficios.

El área de la lámina foliar, la capacidad fotosintética y otras características de las hojas son de gran importancia para determinar la resiliencia de las especies vegetales ante las condiciones climáticas (Niinemets et al. 2001, Fyllas et al. 2009, Malhado et al. 2009). De manera similar, el periódo de floración, el periódo de fructificacion, y el estadio de desarrollo de los organismos son utilizado para estudiar los ciclos de vida de las especies vegetales, su capacidad de adaptación ante cambios ambientales y su riesgo de mortalidad (Wright et al. 2011). Sin embargo, la recopialción de esta información requiere de mucho trabajo de campo y largos periodos de muestreo. Más aún, si se desea comparar diferentes sitios o taxa, es necesario revisar la literatura cientifica y tratar de rescatar la información de los artículos. Sin embargo, pocas publicaciones suelen incluir los "datos crudos". De modo que es difícil para un solo investigador recolectar suficiente información para analizar de manera compresiva un taxon. Con el advenimiento de las tecnologías de la comunicación y la Internet, esto se ha vuelto más sencillo. Los investigadores pueden ahora compartir toda su información de manera transparente y nuevos repositorios son creados para almacenar dicha informacion. El problema actual no radica en el

acceso a la información, sino en como analizar las grandes cantidades que estamos empezando a acumular. De allí la importancia de la ciencia de datos en el campo de la biología.

Este proyecto consiste en la exploración, limpieza, procesamiento y análisis de una base de datos del género *Dipteryx* en América. Se utilizan diferentes tecnicas aprendidas en el curso, con el fin de obtener información sobre el area de la lámina foliar y la radiación solar a la que estan expuestas los árboles.

Objetivo

Realizar un análisis exploratorio de datos fenotípicos de especies del género *Dipteryx* en relación con variables ambientales.

Materiales y Métodos

Los datos utilizados corresponden a informacion descargada de TRY - Plant Trait Database (Kattge et al., 2020; https://try-db.org/TryWeb/Home.php). En esta base de datos se puede solicitar la descarga de información por especie o por caracter. En este caso, se extrajo toda la informacion disponible para especies del género *Dipteryx*. La información se proporciona en un archivo de texto tabular, donde cada fila corresponde a una entrada, categorizadas como caracteres, covariables, o metadatos. Esto resulta en multiples entradas por registro, *i.e.* varias filas por individuo. Los datos se pueden descargar con el siguiente enlace: https://docs.google.com/spreadsheets/d/1Np844SnnurA6MuLD1lmzmXoS6KGfuNCU/edit?usp=sharing&ouid

Para el analisis de los datos, se cargo el archivo en R utilizando el paquete readxl y se realizó una exploración general con el paquete skimr. Los datos luego fueron filtrados para dejar solo las variables de interés y eliminar datos perdidos. Dada la configuración original de la base de datos, se realizó un reacomodo de los datos. Finalmente, los resultados fueron graficados utilizado ggplot. Para mayor detalle, en la sección de resultados se inclye el código utilizado, asi como los resultados del mismo. Lo anterior solo con carácter demostrativo y para efectos de la evaluación del curso.

Resultados

A continuación se presenta los principales resultados de este análisis exploratorio.

```
#Cargar datos
library(readxl)
data <- read_excel("Dipteryx.xlsx", sheet = 1)

#Resumen de los datos
```

```
library(skimr)
skim(data)
```

Table 1: Data summary

Name	data
Number of rows	3773
Number of columns	28
Column type frequency:	
character	17
numeric	11
Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
LastName	0	1.00	5	10	0	15	0
FirstName	0	1.00	3	13	0	16	0
Dataset	0	1.00	16	59	0	18	0
SpeciesName	0	1.00	14	31	0	11	0
AccSpeciesName	0	1.00	14	18	0	6	0
TraitName	3583	0.05	23	108	0	10	0
DataName	0	1.00	5	142	0	186	0
OriglName	0	1.00	1	74	0	272	0
OrigValueStr	6	1.00	1	99	0	953	0
OrigUnitStr	1893	0.50	1	12	0	41	0
ValueKindName	3504	0.07	4	18	0	4	0
OrigUncertaintyStr	3768	0.00	9	10	0	2	0
UncertaintyName	3768	0.00	11	11	0	1	0
UnitName	3512	0.07	1	8	0	3	0
Reference	0	1.00	6	1774	0	18	0
Comment	1281	0.66	5	578	0	155	0
StdValueStr	3705	0.02	6	19	0	26	0

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
DatasetID	0	1.00	364.76	122.23	28.00	412.00	412.00	412.00	681.00	
AccSpeciesID	0	1.00	18800.77	1.21	18796.00	18801.00	18801.00	18801.00	18804.00	
ObservationID	0	1.00	2712177.87	876887.62	221146.00	2758182.00	3032485.00	3067894.00	4066292.00	
ObsDataID	0	1.00	27369098.08	8830400.37	2241022.00	27940543.00	30510932.00	30910209.00	40535146.00	
TraitID	3583	0.05	3112.57	5.29	3086.00	3106.00	3115.00	3117.00	3117.00	
DataID	0	1.00	3801.02	3100.21	19.00	469.00	2988.00	7026.00	7316.00	
Replicates	3750	0.01	1.96	1.72	1.00	1.00	1.00	2.50	6.00	
StdValue	3264	0.13	2502.80	17248.16	-95.11	-14.72	10.43	32.52	191315.00	
RelUncertaintyPercent	3768	0.00	66.00	0.00	66.00	66.00	66.00	66.00	66.00	
OrigObsDataID	3758	0.00	11580574.73	4342183.28	3241092.00	8824133.00	8982890.00	8945658.00	40182178.00	
ErrorRisk	3583	0.05	2.22	1.09	0.51	1.34	2.08	3.10	5.18	

El set de datos consta de 28 variables (columnas) y 3773 entradas (filas). Para algunas variables se tiene un alto porcentaje de datos perdidos. Estas variables corresponden tanto a clasificadores específicos que utilza la base de datos TRY, como a metadatos o métricas asociadas que no estan disponibles o no es posible estimar para este tipo de caracter. Sin embargo, su ausencia es inconsecuente para análisis posteriores. Los datos de interés se almacenan principalmente en la columna "OrigValueStr", la cual tiene un 99% de completitud (solo 6 valores perdidos). Sin embargo, se evidencia que mucha de la información en otras columnas no es relevante, por ejemplo: los nombres de la persona que sometió la información a la base de datos, o los códigos internos que utiliza TRY para clasificar los datos. Por lo tanto se procede a filtrar los datos, creando un subset solo con las variables de interés.

```
#Crear subset
library(dplyr)
```

```
Attaching package: 'dplyr'

The following objects are masked from 'package:stats':

    filter, lag

The following objects are masked from 'package:base':

    intersect, setdiff, setequal, union
```

```
data %>% select(SpeciesName, ObservationID, DataName, OriglName,    OrigValueStr,   OrigUnitS
#Eliminar NA
data <- data %>% filter(!is.na(OrigValueStr))
```

El set de datos resultante contiene solo 6 columnas y 3767 entradas. Se han eliminado los valores perdidos. Despues de filtrar los datos, es necesario identificar con cuantas especies estamos trabajando.

```
#Cuantas especies?
unique(data$SpeciesName)
```

```
 [1] "Dipteryx panamensis"             "Dipteryx alata"
 [3] "Dipteryx oleifera Benth."        "Dipteryx oleifera"
 [5] "Dipteryx micrantha"              "Dipteryx magnifica"
 [7] "Dipteryx odorata"                "Dipteryx punctata"
 [9] "Dipteryx odorata (Aubl.) Willd." "Dipteryx punctata (Blake) Amsh."
[11] "Coumarouna odorata"
```

En el caso de *D. odorata* y *D. punctata* los nombres aparecen escritos de diferentes formas, lo cual podria generar errores, pues se trata de la misma especie. Se necesita reemplazar los nombres, para uniformizar las entradas.

```
#Establecer condiciones de reemplazo
original <- c("Dipteryx oleifera Benth.", "Dipteryx odorata (Aubl.) Willd.",
              "Dipteryx punctata (Blake) Amsh.", "Coumarouna odorata")
reemplazo <- c("Dipteryx oleifera", "Dipteryx odorata", "Dipteryx punctata",
               "Dipteryx oleifera")
#Reemplazar valores
data$SpeciesName <- replace(data$SpeciesName, data$SpeciesName %in% original,
                            reemplazo)
```

```
Warning in x[list] <- values: number of items to replace is not a multiple of
replacement length
```

```
#Valores unicos
unique(data$SpeciesName)
```

```
[1] "Dipteryx panamensis" "Dipteryx alata"      "Dipteryx oleifera"
[4] "Dipteryx odorata"    "Dipteryx punctata"   "Dipteryx micrantha"
[7] "Dipteryx magnifica"
```

Ahora se cuenta con una lista homogenea de especies.

```
#Mostrar primeras lineas
head(data)
```

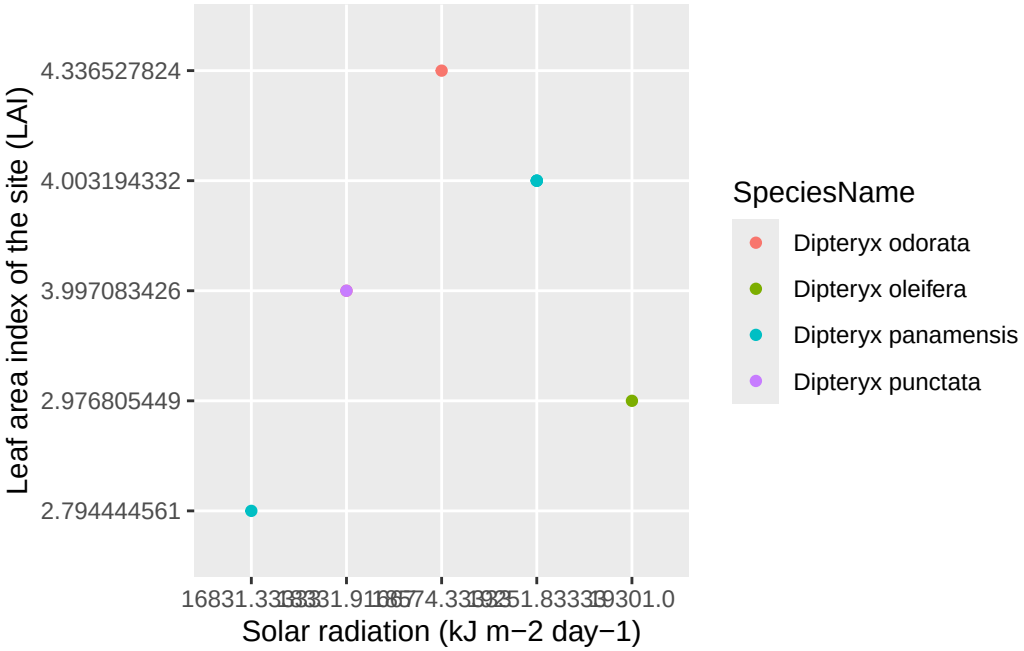
```
# A tibble: 6 x 6
  SpeciesName         ObservationID DataName  OriglName OrigValueStr OrigUnitStr
  <chr>                       <dbl> <chr>     <chr>     <chr>        <chr>
1 Dipteryx panamensis        221146 Plant de~ Seedling~ T            <NA>
2 Dipteryx panamensis        221146 Location~ Geography 10°46'N, 84~ <NA>
3 Dipteryx panamensis        221146 Latitude  Latitude  10.77        dec
4 Dipteryx panamensis        221146 Longitude Longitude -84.03       dec
5 Dipteryx panamensis        221146 Vegetati~ Communit~ tropical mo~ <NA>
6 Dipteryx panamensis        221146 Mean dai~ Daily PP~ 0.38         mol/m2/day
```

Otro problema que se encuentra en este set de datos, es que las variables realmente estan en las filas, mientras las columnas solo representan categorias del tipo de datos. Por esta razon, existen multiples filas o entradas para el mismo individuo. Es necesario extraer solo las variables de interés, en este caso: área foliar y radiación solar de las muestras que tengan información para ambas.

```
#Filtrar entradas con valores de radiacion solar
radiacion <- data[data$DataName == "Solar radiation (kJ m-2 day-1)",]
individuos_radiacion <- unique(radiacion$ObservationID)

#Filtrar entradas con valores de area foliar
area <- data[data$DataName == "Leaf area index of the site (LAI)",]
individuos_area <- unique(area$ObservationID)

#Intersección de individuos que tienen ambos registros
individuos_ambos <- intersect(individuos_radiacion, individuos_area)

#Filtrar datos solo de individuos en comun
radiacion <- radiacion[radiacion$ObservationID %in% individuos_ambos, ]
area <- area[area$ObservationID %in% individuos_ambos, ]
```

Ahora es posible graficar los datos de área foliar contra los valores de radiación solar.

```
library(ggplot2)

#Grafico de SRAD vs LAI
ggplot(data = radiacion, aes(x = OrigValueStr, y = area$OrigValueStr, color = SpeciesName)) +
```



Como se puede observar, no hay una relación lineal entre el área foliar y los niveles de radiación solar. El máximo de área se alcanza en los sitios con *D. odorata*, pero la mayor radiación se reporta en los sitios con *D. oleifera*. Se requiere de un estudio más detallado y con un tamaño de muestra mayor para dilucidar las relaciones entre área foliar y radiación solar. Asi como modelos estadísticos más robustos.

Discusión

La tabulación de la base de datos reflejó un arreglo más difícil de solucionar, apesar de hacer la exploración y limpiar los datos. Al estar las variables realmente estaban almacenadas en las filas y no en columnas, se dificulta poder usar la información. Más aún no hay un patrón claro que permita ordenar los datos (*i.e.* transponer las filas para generar columnas con las variables), pues para cada registro el numero de filas (variables) era diferente dependiendo de la fuente de la información.

Según el gráfico de dispersión de las variables radiación solar e índice de área foliar, no parece existir una relación clara entra las variables. Aunque se cumple parcialmente la expectativa de observar una disminución en el área foliar conforme incrementa la radiación solar. Pese a que el set de datos cuenta con 147 observaciones independientes (individuos diferentes), muy pocos cuentan con información completa para todas las variables. Por ejemplo revisando las variables presentes en la columna DataName nos damos cuenta que existen pocas variables relacionadas a datos fenologicos y las que existen son de muy poco individuos.

En conclusion, el formato de la base de datos es crucial para su utilización. Aunque TRY proporciona un recurso muy valioso, es necesario que se defina un estandar para todos los investigadores que desean contribuir datos. Pues de lo contrario, se vuelve muy dificil poder extraer la información. La exploración de los datos es vital a la hora de trabajar con bases de datos, para deteminar que tipo de análisis realizar y obtener resultados que contribuyan a la toma de decisiones. Debido a lo que se observó en la exploración de los datos se tuvo que hacer filtros para poder trabajar con las variables, sin embargo la base de dato requiere de mucha depuración.

Referencias

Kattge, J., G. Bönisch, S. Díaz, S. et al. (2020) TRY plant trait database – enhanced coverage and open access. Global Change Biology, 26:119–188.

Niinemets, U. (2001). Global-scale climatic controls of leaf dry mass per area, density, and thickness in trees and shrubs. Ecology 82:453-469.

Fyllas, N. M., S. Patino, T. R. Baker, et al. (2009). Basin-wide variations in foliar properties of Amazonian forest: phylogeny, soils and climate. Biogeosciences 6:2677-2708.

Malhado ACM, Malhi Y, Whittaker RJ, et al. (2009). Spatial trends in leaf size of Amazonian rainforest trees. Biogeoscience 6, 1563-1576.

Wright, S. J., K. Kitajima, N. J. B. Kraft, et al. (2011). Functional traits and the growth-mortality tradeoff in tropical trees. Ecology 91:3664-3674.